Describing and Exploring Data

Without statistics we are cast adrift on an ocean of confusion, but armed with stats we can take control of our lives, hold our rulers to account and see the world as it really is. What's more, Hans concludes, we can now collect and analyse such huge quantities of data and at such speeds that scientific method itself seems to be changing.

Hans Rosling

http://www.open.edu/openlearn/whats-on/tv/the-joy-stats

http://flowingdata.com/2010/11/30/the-joy-of-stats-with-hans-rosling/

"Remember, statistics are in the eye of the manipulator."

Statistics means never having to say you are certain

Types of Statistics

- Descriptive statistics
- Inferential Statistics
- Univariate Statistics
- Bivariate Statistics
- Multivariate Statistics

Descriptive Statistics

Reminder

- Nominal or Ordinal – Categorical variables
- Interval or Ratio – Numerical Variables

Univariate Analyses

- Categorical Data
  - Frequencies
  - Charts
- Numerical data
  - Descriptive measures

Frequency Tables

Usually categorical data, or interval or ratio data classified into intervals

Go to Example Data –

construct frequency tables for

Q12.5, Q12.6 & Q17

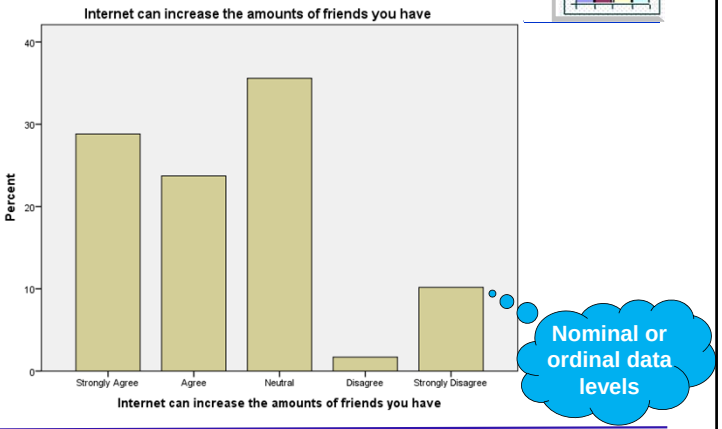
Corrie Uys
UysC@cput.ac.za
Cape Peninsula University of Technology
Centre for Postgraduate Studies

Internet can increase the amounts of friends you have

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly Agree	34	27.0	28.8	28.8
	Agree	28	22.2	23.7	*52.5*
	Neutral	42	33.3	35.6	*88.1*
	Disagree	2	1.6	1.7	*89.8*
	Strongly Disagree	12	9.5	10.2	*100.0*
	Total	118	*93.7*	*100.0*	
Missing	System	8	6.3		
Total		126	*100.0*		

Corrie Uys
UysC@cput.ac.za
Cape Peninsula University of Technology
Centre for Postgraduate Studies

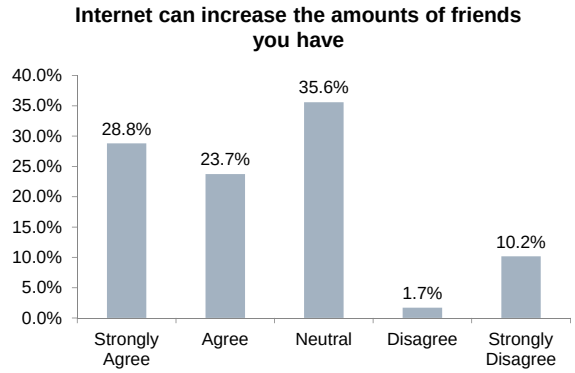
Bar charts



Nominal or ordinal data levels

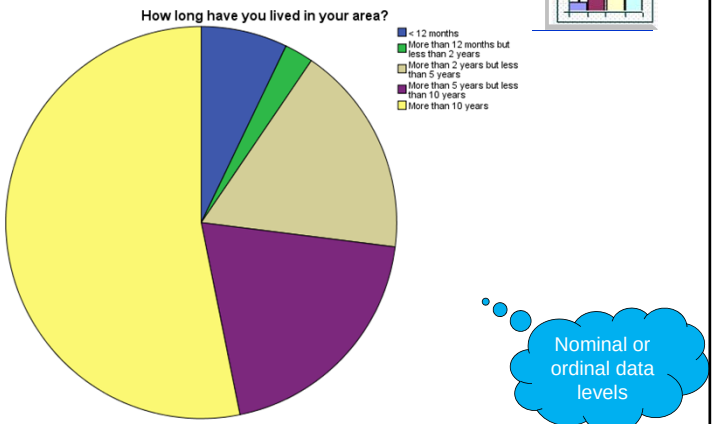
Corrie Uys
UysC@cput.ac.za
Cape Peninsula University of Technology
Centre for Postgraduate Studies

Bar charts from Excel



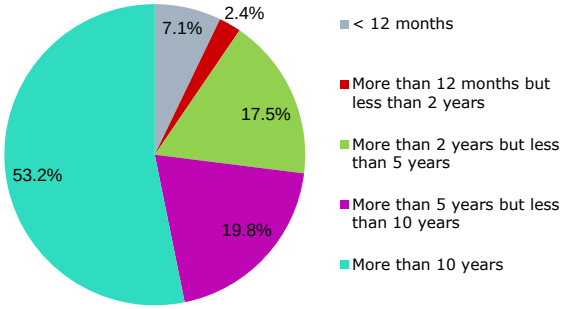
Corrie Uys
UysC@cput.ac.za
Cape Peninsula University of Technology
Centre for Postgraduate Studies

Pie charts



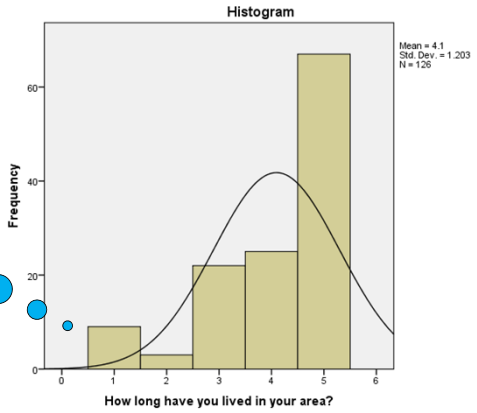
Nominal or ordinal data levels

Pie charts from Excel

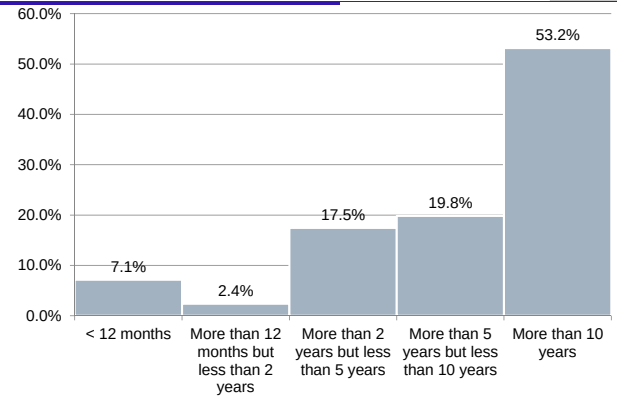


Histograms

Usually interval or ratio data; classified into intervals



Histogram using Excel



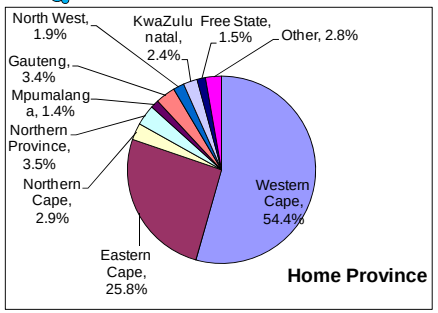
Cape Peninsula University of Technology Centre for Postgraduate Studies
Corrie Uys
UysC@cput.ac.za

Pie Charts

Presenting percentages

Pie Chart

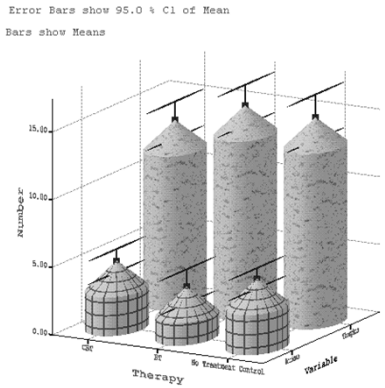
Home Province	Frequency	Percentage
Western Cape	1583	54.4%
Eastern Cape	751	25.8%
Northern Cape	84	2.9%
Northern Province	103	3.5%
Mpumalanga	41	1.4%
Gauteng	99	3.4%



Cape Peninsula University of Technology Centre for Postgraduate Studies
Corrie Uys
UysC@cput.ac.za

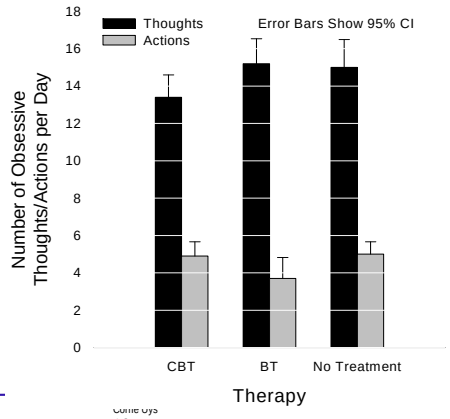
Why is this Graph Bad?

- 3-D effect – obscures data
- Patterns – distracting
- Cylindrical Bars – distracting
- Badly labelled y-axis



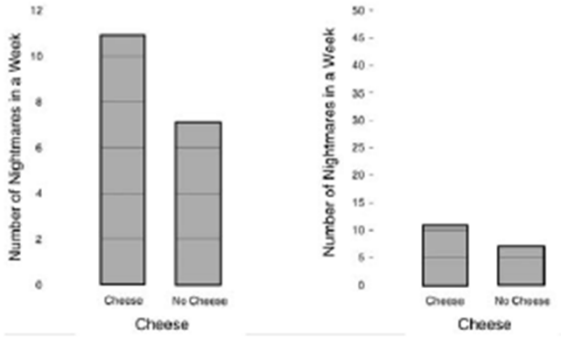
Cape Peninsula University of Technology Centre for Postgraduate Studies
UysC@cput.ac.za

Why is this Graph Better?



Cape Peninsula University of Technology Centre for Postgraduate Studies
UysC@cput.ac.za

Deceiving the Reader



The Art of Presenting Data

- Graphs should (Tufte, 2001):
  - Show the data.
  - Induce the reader to think about the data being presented (rather than some other aspect of the graph).
  - Avoid distorting the data.
  - Present many numbers with minimum ink.
  - Make large data sets (assuming you have one) coherent.
  - Encourage the reader to compare different pieces of data.
  - Reveal data.

Describing Data

Numerical Data

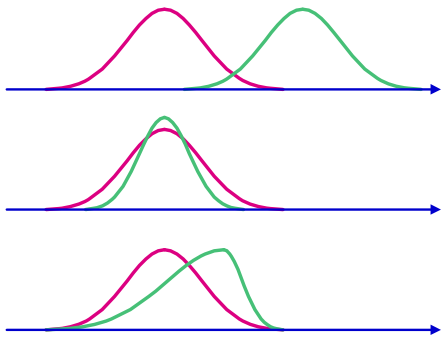
Descriptive Measures

Numerical Data Properties

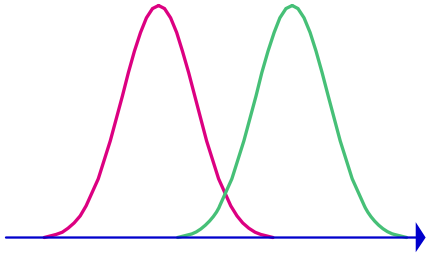
Central Tendency (Location)

Variation (Spread)

Shape



Measures of Central Tendency



Mean

Only interval or ratio data levels

- Arithmetic ***Average***
- Most common measure of central tendency
- Affected by extreme values ('outliers')
  - Population mean – μ ("mu")
  - Sample mean – $\bar{x}$
- Example: Age: $\bar{x}$ = 24.26

$$= \frac{Sum\ of\ all\ values}{Number(count)\ of\ values}$$

Age
17
34
35
36
18
22
27
28
24
21
19
23
18
24
26
21
17
22
29

Median

All levels of data

- Measure of central tendency
- Middle value in ordered sequence
  - If odd *n*, middle value of sequence
  - If even *n*, average of 2 middle values
- Not affected by extreme values
- Position of median in sequence

Age	Age
17	17
34	17
35	18
36	18
18	19
22	21
27	21
28	22
24	22
21	23
19	24
23	24
18	26
24	27
26	28
21	29
17	34
22	35
29	36

Mode

All levels of data

- Measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- May be no mode or several modes
- May be used for numerical & categorical data

Age
17
34
35
36
18
22
27
28
24
21
19
23
18
24
26
21
17
22
29

Age
17
17
18
18
19
21
21
22
22
23
24
24
26
27
28
29
34
35
36

More measures of location - Quartiles

Age				
17	5.3%			
17	10.5%	Q1	First Quartile	18.75
18	15.8%			
18	21.1%			
19	26.3%			
21	31.6%			
21	36.8%	Q2	Second Quartile = Median	22.5
22	42.1%			
22	47.4%			
23	52.6%			
24	57.9%			
24	63.2%	Q3	Third Quartile	27.3
26	68.4%			
27	73.7%			
28	78.9%			
29	84.2%			
34	89.5%	Q4	Fourth Quartile = maximum	36
35	94.7%			
36	100.0%			

Exercise

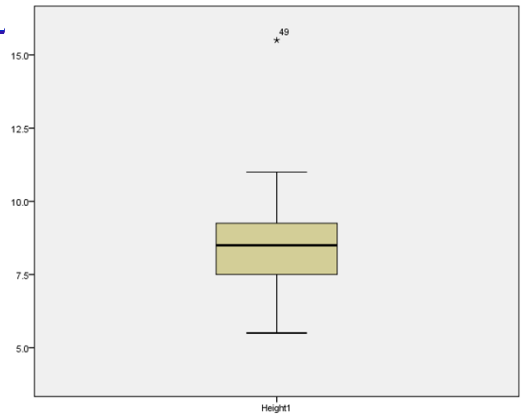
- Import PlantGrowth.xlsx into SPSS 22
- Find means, medians, etc

Boxplots (Box-Whisker Diagrams)

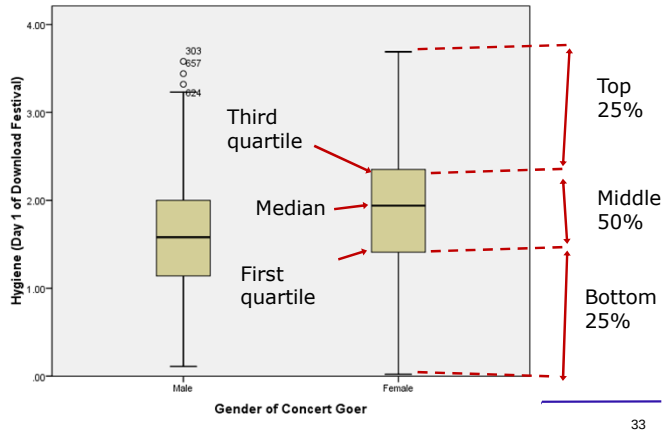
- Boxplots are made up of a box and two whiskers.
- The box shows:
  - The median
  - The third and first quartile
  - The limits within which the middle 50% of scores lie.
- The whiskers show
  - The range of scores
  - The limits within which the top and bottom 25% of scores lie

Usually interval or ratio data

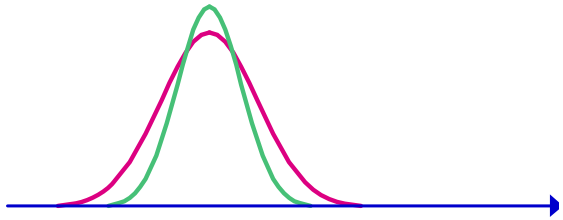
Boxplots (Box-Whisker Diagrams)



What Does The Boxplot Show?



Measures of Spread



Range (or Midrange)

- Distance between smallest value and largest value
  - Minimum = 17
  - Maximum = 36
  - Range = 36 – 17 = 19

Standard Deviation and Variance

- Shows how data are distributed
- Show variation about mean $\bar{x}$ or μ
  - Population standard deviation = σ
  - Sample standard deviation = s
  - Population variance = σ^2
  - Sample variance = s^2

Standard deviation

Only interval or ratio data levels

- Distance measure
- "Average (standardised)" distance of all the values to the mean
- Here standard deviation (s) = 5.961
- *A rough estimate of the standard deviation is the range/4.*

Age
17
34
35
36
18
22
27
28
24
21
19
23
18
24
26
21
17
22
29

The Standard Deviation and the Shape of a Distribution

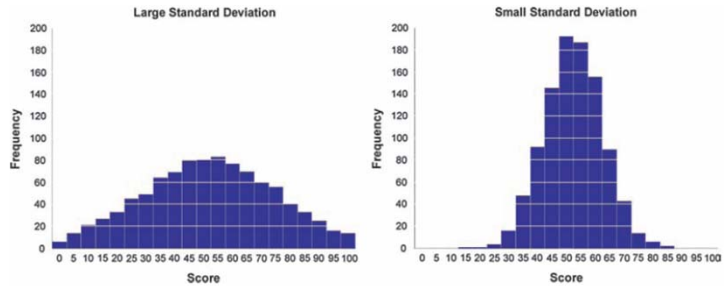


FIGURE 2.6 Two distributions with the same mean, but large and small standard deviations

Variance

Only interval or ratio data levels

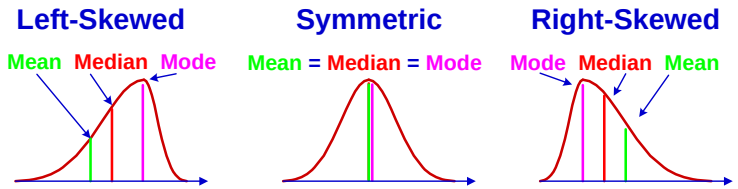
- Square of the standard deviation.
- Used in the formula of most statistical calculations when proving hypotheses
- Here variance (S^2) = 35.538

Age
17
34
35
36
18
22
27
28
24
21
19
23
18
24
26
21
17
22
29

Shape

Ordinal, interval or ratio data levels

- Describes how data are distributed
- Measures of shape
  - Skew: Symmetry



Descriptive Statistics
Mean/average and Variability

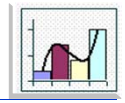
Use PlantGrowth SPSS data to find measures of spread

Bivariate analyses

- Crosstabulation Tables / Contingency Tables
  - Stacked/Clustered Bar Charts
- Correlation analyses
  - Scatter Plots
  - Line Plots
- Linear Regression
  - Line Plots
- Comparing Means/Medians of two groups

Years of Studies Have Proved
Infertility Less Likely in Women With Children
New York Times headline

To read…



- Writing an academic journal article, by Dr Theuns Kotze
  web.up.ac.za/sitefiles/file/40/753/writing_an_academic_journal_article.pdf

Cape Peninsula University of Technology
Centre for Postgraduate Studies

Corrie Uys
UysC@cput.ac.za